DOCUMENT RESUME

ED 309 193 TM 013 673

AUTHOR Kogut, Jan
TITLE Review of IRT-Based Indices for Detecting and Diagnosing Aberrant Response Patterns. Project Psychometric Aspects of Item Banking No. 10. Rapport 86-4.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE 86
NOTE 42p.
AVAILABLE FROM Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algorithms; Computer Simulation; Foreign Countries; *Latent Trait Theory; Statistical Analysis; Testing Problems; *Test Items
IDENTIFIERS *Aberrance; Extended Caution Index; Indicators; Rasch Model; Residuals (Statistics); *Response Patterns

ABSTRACT

Methods and indices based on item response theory (IRT) for detecting and diagnosing aberrant response patterns are reviewed. These indices are divided into three groups: (1) residuals-based; (2) likelihood-based; and (3) ratio of covariances-based (extended cautions). For each index, the determination of its sampling distribution as well as its critical value to declare aberrance (with a known Type I error) are considered as the central questions. Applications of the indices in practical testing situations and in simulation studies are briefly described. All indices are discussed from the perspective of testing certain hypotheses about a person's ability. A comparison between indices is provided for the Rasch model in particular. Two of the indices are suggested for use in a Test Service System. A procedure and algorithm are provided to detect aberrants in tests in which the items are already of Rasch's structure. One graph illustrates an expected and observed person response curve. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED309193

Review of IRT-based Indices for Detecting and Diagnosing Aberrant Response Patterns

Rapport 86-4

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

J. Kogut

Vakgroep
Onderwijskundige Meetmethoden
en Data-analyse

TM 013073

Toegepaste Onderwijskunde

Universiteit
Twente

Project Psychometrische Aspecten van Item Banking No. 10

Colophon

Typing : L. Padberg
Cover design : M. Driessen, AV-section, University of Twente
Printed by : Central Reproduction Department, University of Twente

A Review of IRT-Based Indices for Detecting and Diagnosing Aberrant Response Patterns

Jan Kogut

Abstract

In this paper, methods and indices based on item response theory for detecting and diagnosing aberrant response patterns are reviewed. These indices are divided into three groups: residuals-, likelihood- and ratio of covariances-based indices. For each index, the determination of its sampling distribution as well as its critical value to declare aberrance (with a known type I error) are considered as the central questions. Applications of indices in practical testing situations and in simulation studies are described shortly. All indices are discussed from the point of view of testing certain hypotheses about a person's ability and in particular, for the Rasch model, a comparison between indices is provided. Finally, some indices and procedures to detect aberrant response patterns and to diagnose the specific kind of their aberrance for use in a Test Service System are suggested.

A Review of IRT-Based Indices for Detecting and Diagnosing Abberant Response Patterns

Introduction

In every day test-taking practice, even when a specific item response theory (IRT) model is valid for most persons, many disturbances can occur that affect the item-response behavior of certain persons. Some examples of these disturbances are: copying, cheating, a slow startup, guessing to complete, and cultural or educational bias. The observed response pattern, then, consists of relatively many responses that are very unexpected under the IRT model used. Such a pattern typically looks like one in which a low-scoring student gets relatively many hard items correct or a pattern in which high-scoring student misses relatively many easy items. Owners of such a pattern and their response patterns are called aberrant, unusual or atypical. If there is a large difference between person's observed and expected patterns, the usual IRT ability estimate as a measure of a person's true ability is not acceptable, or at least, less reliable. Therefore, aberrant persons should be detected and treated separately. They can be removed from the data in order to get better item and person parameter estimates. Further, diagnoses of the kind of observed aberrance may lead to a more accurate assessment of the individual

or subgroup performance on the test. In particular, it may be possible to identify subgroups of response patterns of a certain kind of aberrance (e.g., the presence of cheaters or students unfamilar with the specific subdomain). In turn, this may lead to more appropriate actions in future test administrations or in the instructional process (e.g., against cheaters, or to improve teaching). This is why it is recommended to make diagnoses of the kind of detected aberrance as carefully as possible.

In the last decade, a variety of indices for detecting aberrant response patterns have been developed in the IRT literature. Some of these indices have been proposed specifically for the Rasch model (Molenaar, 1986; Smith, 1985; Wright, 1977; Wright & Stone, 1979, chap.4, chap.7), others for more general models (Drasgow, Levine & Williams, 1985; Levine & Drasgow, 1983; Levine & Rubin, 1979; Tatsuoka, 1984; Trabin & Weiss, 1983; Wright & Masters, 1982, chap.5). Several authors have performed comparisons of the indices in order to determine a relationship between them (e.g., Birenbaum, 1985; Harnisch & Tatsuoka, 1983), to evaluate their relative usefulness in detecting aberrants (e.g., Levine & Drasgow, 1983; Levine & Rubin, 1979; Rudner, 1983), and to study their robustness against the variability in item and person parameters (Hoijtink, 1986a).

As a matter of fact, all indices measure the degree of consistency of an observed pattern to the IRT model used. Most indices measure the overall person's fit to the model, and can thus be used to identify aberrants. Others, however, measure a more

specific kind of person's fit (Drasgow & Levine, 1983; Smith, 1985, 1986; Trabin & Weiss, 1983); and can be applied for computerized diagnoses of the kind of aberrance. Extended caution indices measure the similarity of a person's pattern to the group- or individual-normed one (Tatsuoka, 1984; Tatsuoka & Linn, 1983); these are helpful in student-class comparisons and to diagnose students who consistently use an erroneous rule of operation. Unfortunately, the percentiles and distribution of all these indices (in the null situation of no aberrant person) are not sufficiently known yet (Hoijtink, 1986a). In account of that, the probability of misclassifying an IRT-behaving person as aberrant (i.e., of the type I error) is still hardly recognized. Little is known about their robustness in the presence of aberrants in the sample (Levine & Drasgow, 1983), nor about their power (Drasgow & Levine, 1986; Levine & Drasgow, 1984). However, quite recently some progress has been made for the case of the Rasch model (Hoijtink, 1986a,b; Molenaar, 1986).

In this paper, the one-, two- and three-parameter logistic model are used. If P_{vi} denotes the probability of the correct response given by person v (with ability θ) to item i, then the three-parameter logistic (3PL) model is written as

$$P_{vi}(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_v - b_i)]}, \qquad \begin{matrix} v = 1,\ldots,n; \\ i = 1,\ldots,L; \end{matrix}$$

where $a_i > 0$, $b_i \in R$ and $0 \leq c_i < 1$ are the parameters

characterizing item i; the restriction $c_i = 0$ for all $i = 1,\ldots,L$ results in the two-parameter logistic (2PL) model, whereas $c_i = 0$ and $a_i = 1$ for all $i = 1,\ldots,L$ gives the one-parameter logistic (1PL) model, also called the Rasch model (RM). All indices reviewed in this paper are divided into three groups in regard to the elements they are built of. The residual-, likelihood- and the ratio of covariances (extended caution) -based indices are reviewed in the following sections. Next, all indices are discussed from the point of view of testing certain hypotheses about a person's ability and a comparison between indices for the RM is provided. Finally, some indices and procedures for use in a Test Service System (TSS) are suggested.

Residual-Based Indices

IRT gives us the possibility to account the probability of a correct response on an item for a person. Hence, the expected score on an item or a subset of items can be calculated. By substracting the expected score from an observed one, a residual is obtained. Accumulating residuals over a person's responses provide us with a measure of the person's fit to the model, a fit statistic. A significantly large value of the fit statistic is an indication of nonfit (or aberrance) of that person or his/her response pattern. So, several authors have proposed residual-based indices for the RM (Smith, 1985; Wright, 1977; Wright & Stone, 1979, chap.4, chap.7), others for more general models (Trabin & Weiss, 1983; Wright &

Masters, 1982, chap.5).

Wright and his associates have defined two indices of the overall person's fit for the RM model (Wright, 1977; Wright & Stone, 1979, chap.4, chap.7). The first, UT, the unweighted total fit statistic of the given person v, is defined as

$$UT = \frac{1}{(L-1)} \sum_{i=1}^{L} \frac{(x_{vi}-P_{vi})^2}{P_{vi}(1-P_{vi})},$$

where x_{vi} is the observed response and P_{vi} is the probability of the correct response of person v to item i. The second index, WT, the weighted total fit statistic, is defined by

$$WT = \frac{\sum_{i=1}^{L} (x_{vi}-P_{vi})^2}{\sum_{i=1}^{L} P_{vi}(1-P_{vi})}.$$

Note that the expected value and variance of x_{vi} are $E(x_{vi}) = P_{vi}$ and $Var(x_{vi}) = P_{vi}(1-P_{vi})$. Thus, UT is a corrected mean of squared standardized residuals, summed over all items for the given person. These standardized residuals are weighted in WT by their variances. The unweighted total fit statistic UT is rather sensitive to the event of just one unexpected response on an item with a difficulty far away from the ability of that person. In contrast, the weighted total fit statistic WT is much less influenced by such particular unexpected responses. Both indices, UT and WT, measure the overall fit of a person to the model, and a relatively large value of the

indices identifies that person as aberrant. A critical value to determine aberrance might be set equal to the $(1-\alpha)^*$ 100-th percentile of the index distribution in the null situation (i.e., no aberrant person). By this critical value, the probability of a misclassification of an IRT-behaving person as aberrant (the type I error) will be α. However, there are some questions to be answered previously. First, what is the distribution of these indices in the null situation? Second, usually the P_{vi} are unknown and have to be estimated from the data through an estimation of item and person parameters. How will the indices be affected by these estimates? Third, when computing the indices, aberrants are still unidentified. How will the presence of aberrant persons affect the estimation of parameters and, as a consequence, affect the indices? These problems are still unsatisfactorily solved; however, some progress (reported below and in the next sections) has been made.

Wright and his associates (Wright & Stone, 1979, chap.4, chap.7; Wright & Masters, 1982, chap.5) have claimed that the standardized transformations of the UT and WT indices, namely,

$$ZUT = \{\ln(UT+UT-1\} \{(L-1)/8\}^{1/2}$$

for UT and

$$ZWT = (WT^{1/3}-1)(3/q) + (q/3)$$

for WT (where q is the variance of WT) have approximately the

standard normal distribution. Therefore, a fixed percentile of the standard normal distribution has been used as the critical value to indicate aberrants. However, Hoijtink (1986a) has shown in a simulation study, for the RM (i.e. in the null situation), that distributions of ZUT and ZWT are rather far from the normal distribution. In addition, a fixed percentile of the index distribution differs as a function of both item difficulty and ability distributions as well as per test score group. As a consequence, the use of a fixed sampled overall percentile of ZUT as the critical value to indicate aberrants implies that too many RM-behaving persons of extreme ability will be classified as aberrant and too many aberrant persons of medium ability will remain undetected. For the transformation ZWT, reverse conclusions can be drawn. Resuming, to obtain a constant probability of misclassification of an RM-behaving person as aberrant, a separate critical value should be assigned to each test score group (equal to a fixed percentile corresponding to that score group).

For aberrant person already identified, a visual inspection of residuals can be utilized to diagnose a specific kind of aberrance. In general, there are two kinds of aberrances: random (for some items large residuals) and systematic (a pattern of residuals). Typical random aberrances are excessive guessing or excessive sleeping. Examples of systematic aberrances are startup, guessing to complete, item type or educational bias, and misconception or erroneous rule of operation. To diagnose these specific kinds of aberrance a priori specification of items order is needed, e.g.,

the order of difficulty of the items, their position in the test, the item type and different conceptual domains. Then, on the basis of plotting and a visual inspection of person's standardized residuals against the specified item order, decisions about the kind of aberrance can be made. For illustrations of detections and diagnoses of aberrant persons, see Wright & Masters (1982, chap.5) and Wright & Stone (1979, chap.7). General strategies for a graphical analysis of RM residuals are discussed by Ludlow (1986).

To avoid the subjectivity of decisions by visual inspection of standardized residuals, two other person's fit indices have been postulated by Smith (1985) for the RM. The first, UB, the unweighted between fit statistic for a given person v, is defined as

$$UB = \frac{1}{(J-1)} \sum_{j=1}^{J} \frac{\left(\sum_{i \in j}^{L_j} x_{vi} - \sum_{i \in j}^{L_j} P_{vi}\right)^2}{\sum_{i \in j}^{L_j} P_{vi}(1-P_{vi})},$$

where J is the number of fixed non-overlapping subsets of items and L_j is the number of items in subset j. The second statistic, UW, the unweighted within-set fit statistics for any subset of L_j items, is defined by

$$UW = \frac{1}{L_j} \sum_{i=1}^{L_j} \frac{(x_{vi}-P_{vi})^2}{P_{vi}(1-P_{vi})},$$

where P_{vi} is based on the subset of L_j items exclusively. The

distributions and critical values to declare aberrance (of known type I error) for these indices are not known (Smith utilizes 2.0 for both indices). Both indices can be used to clarify causes of the overall nonfit of aberrants, detected with the total fit statistics ZUT and ZWT. Such procedures require a priori specification of item subsets and can be interpreted, for the UB index, as a test of the null hypothesis that the person's ability is the same in each group of items. For example, items can be grouped on the basis of item type in order to detect an item-person interaction (Smith, 1985, 1986). Hence, subjective visual inspection of residuals can now be replaced by more objective procedures.

Another residual procedure to detect and diagnose aberrant persons, is to divide items into strata of homogeneous items and to compare the observed and expected proportion correct by means of an ordinary chi-squared statistic.

Trabin and Weiss (1983) have utilized this approach for the 3PL model with strata obtained by grouping the items into levels of similar difficulty. For a given person v, the expected score on subset j (containing L_j items) can be defined as the sum of their P_{vi}. Therefore, a person's expected proportion correct responses on subset L_j can be referred to as

$$P_{Ej} = \frac{1}{L_j} \sum_{i=1}^{L_j} P_{vi}.$$

The so-called person response curve (PRC) is obtained by a plot of

the within-stratum expected proportion correct against difficulty levels. The PRC has some important properties that are comparable to those of the item characteristic curve (ICC). In fact, it is a monotonically decreasing function asymptotically approaching 1 at the left and 0 at the right side of the item difficulty continuum. In addition, it reaches 0.5 at the difficulty level equal to the person's ability. The steepness of the PRC at this point can be interpreted as the person's precision with respect to the test. In order to obtain the observed PRC, the observed within-stratum proportion correct has to be plotted against difficulty levels (see Figure 1). If a person answers items in accordance with an IRT model, the observed PRC ought to conform to the expected PRC. This is why detection and diagnosis of a person's aberrance is possible.

Insert Figure 1 about here

To determine whether the observed PRC is significantly different from the expected PRC, a chi-square test statistic can be used. That is,

$$\chi^2 = \sum_{j=1}^{J} \frac{(P_{Oj}-P_{Ej})^2}{P_{Ej}} ,$$

where P_{Oj} is the observed, P_{Ej} the expected within-stratum proportion correct, and J the number of strata.

Trabin and Weiss (1983) have claimed that the χ^2 index has approximately the chi-square distribution with J-m-1 degrees of freedom, where m is the number of estimated parameters. In their study, this seems to be the case, but at least some other simulation studies in the null situation should be run before the chi-square distribution of their index might be established. When testing for the overall person's fit, dividing items into many strata and using the chi-square approximation restricts the application of the χ^2 index to long tests only. However, if a small number of strata is chosen, the χ^2 index can be useful to test more specific kinds of person's fit (in a similar way as for the UB index).

In conclusion, for each of the logistic models, all residual-based indices can be calculated with the P_{vi} replaced by their estimates from these models. The residual-based indices provide the most detailed information about a person's deviation from the IRT model. Aberrants can be identified with the help of UT/ZUT or WT/ZWT -- a procedure that can be interpreted as a test of overall goodness of person's fit. Besides, diagnoses of the specific kind of aberrance are possible with the UB or χ^2 indices. This procedure can be interpreted as a test of the null hypothesis that the person's ability is the same in each group of items. Testing this hypothesis is of importance when diagnosising a specific item-person interaction or bias. Different hypotheses about the specific kind of aberrance can be formalized and tested with different specifications of item subsets (by UB) or strata (by χ^2).

Likelihood-Based Indices

In another approach, a person is classified as aberrant if his/her pattern has a very low probability to be generated by an IRT-behaving person of the same ability. For a person detected as aberrant, the kind of aberrance can be diagnosed by fitting a more general IRT model, in which this specific kind of aberrance is modelled as a special case. Levine and his associates have utilized this approach to detect and diagnose aberrants and introduced indices of a new type. They have extensively treated these indices for the 3PL model , both on real and simulated data (e.g., Drasgow, 1982; Drasgow, Levine & Williams, 1985; Levine & Drasgow, 1983; Levine & Rubin, 1979).

The first likelihood-based index, L_0, has been defined as the maximum of the logarithm of the conditional likelihood for the given pattern (Levine & Drasgow, 1983; Levine & Rubin, 1979), i.e.,

$$L_0 = \max_{\theta} \log P(\underline{X}|\theta) = \log P(\underline{X}|\hat{\theta}) ,$$

where

$$P(\underline{X}|\hat{\theta}) = \prod_{i=1}^{L} P_{vi}(\hat{\theta})^{x_{vi}} \left[1-P_{vi}(\hat{\theta})\right]^{1-x_{vi}} ,$$

$\underline{X} = (x_{v1}, x_{v2}, \ldots, x_{vL})$ is the person's pattern of responses and $\hat{\theta}$ is the maximum-likelihood (ML) estimate of the person's ability.

Actually, L_0 is a by-product of the ML estimation of the ability if an algorithm to maximize the conditional likelihood has been used (e.g., Wingersky, Barton & Lord, 1982). Therefore, L_0 is computationally the simplest index proposed so far. Drasgow, Levine and Williams (1985) have shown empirically that the mean of L_0 depends on levels of $\hat{\theta}$ and have proposed a standardized version, denoted as ZL_0.

$$ZL_0 = \frac{L_0 - E(L_0|\hat{\theta})}{Var^{1/2}(L_0|\hat{\theta})},$$

where $E(L_0|\hat{\theta})$ and $Var(L_0|\hat{\theta})$ are the conditional expected value and variance of L_0, respectively, i.e.,

$$E(L_0|\hat{\theta}) = \sum_{i=1}^{L} P_{vi}(\hat{\theta}) \log P_{vi}(\hat{\theta}) + [1-P_{vi}(\hat{\theta})] \log[1-P_{vi}(\hat{\theta})],$$

and

$$Var(L_0|\hat{\theta}) = \sum_{i=1}^{L} P_{vi}(\hat{\theta})[1-P_{vi}(\hat{\theta})] \{\log P_{vi}(\hat{\theta})/[1-P_{vi}(\hat{\theta})]\}^2.$$

From these definitions it follows that L_0 or ZL_0 indicate whether the person has approached the test as an IRT-behaving person of the same ability would do. Thus, a very low value of L_0 or ZL_0 indicates that a given person's pattern has a very low probability to be generated by an IRT-behaving person and allows us to determine that pattern as aberrant.

As a matter of fact, L_0/ZL_0, just like UT/ZUT and WT/ZWT, the

residual-based indices, measure the overall goodness of person's fit to the IRT model. Namely, L_0 can be rewritten as a sum,

$$L_0 = \sum_{i=1}^{L} x_{vi} \log P_{vi}(\hat{\theta}) + [1-x_{vi}] \log[1-P_{vi}(\hat{\theta})] ,$$

and L_0 is near zero if the observed responses are likely under the IRT model. Thus, actually, all these indices measure the overall goodness of person's fit but can differ in regard to their sensitivity to different item and person parameters and their estimates.

Levine and Drasgow (1983) have shown, in a simulation study of the 3PL model, that values of L_0 are not seriously distorted when the exact item parameters are replaced by their sample estimates, neither from the norming sample (i.e. without aberrants), nor from a sample containing a small number of ability-unrelated aberrants. However, these conclusions are not very surprising because the likelihood function (of which the maximum is L_0) does not change very much for values in the neighborhood of the maximizing argument, particularly if the likelihood is relatively flat. Besides, the presence of ability-unrelated aberrants in the sample must have random effects on item and person estimates, so these effects ought to cancel out. Standard errors of item and persons estimates can possibly increase, but without inducing bias of the estimates. In the case of ability-related kinds of aberrance, the value of L_0 can be really distorted and therefore the seriousness of bias in L_0, as well as in item and person estimates, should be

reevaluated.

Drasgow, Levine and Williams (1985) have argued that for the 3PL model the left tails of the distribution of ZL_0, within different ability intervals, are approximately the same as of the standard normal distribution. Being aware of that, the lower tail percentiles of the standard normal distribution have been employed to classify aberrants.

Most IRT models assume a constant ability during the test. In some types of aberrance, the aberrant person behaves as a person with a fluctuating ability (i.e. varying from item to item during the test). Such a person can be modelled with the help of the so-called variable-ability models. Then, by comparing the results of these two models, the ability-fluctuation aberrance of persons can be detected. Levine and Drasgow (1983) have used this method for the 3PL model in connection with the Gaussian model assuming that for each item ability is independently sampled from the normal distribution $N(\theta_G, \sigma_G^2)$. The anologue of L_0 in the Gaussian model is defined as follows:

$$L_G = \max_{\theta_G,\ \sigma_G^2} \log\ P(\underline{X}\,|\,\theta_G, \sigma_G^2)\ ,$$

where

$$P(\underline{X}\,|\,\theta_G, \sigma_G^2) = \prod_{i=1}^{L} \int P_{vi}(t)^{x_{vi}}\ [1-P_{vi}(t)]^{1-x_{vi}}\ \Phi(t\,|\,\theta_G, \sigma_G^2)dt$$

is the marginal likelinood and $\Phi(t\,|\,\theta_G, \sigma_G^2)$ is the normal density

function. Since the IRT model can be regarded as a submodel of the Gaussian model (with $\sigma_G=0$), an improvement of the fit is expected. To measure the degree of this improvement, two indices have been considered:

$$LR = L_G - L_0$$

and

$\hat{\sigma}^2$ = the ML estimate of the ability variance σ_G^2,

as obtained during the calculation of L_G. If values of these indices are significantly larger than zero, the Gaussian model fits a given person's pattern significantly better than the IRT model and ability-fluctuation aberrance can be diagnosed. Thus, both indices can be employed to test the null hypothesis (ability is constant across items) against the alternative hypothesis (ability is a normal random variable across items). Ability-fluctuation aberrance can be identified when the LR and $\hat{\sigma}^2$ reach relatively large values.

The distribution of LR and $\hat{\sigma}^2$ in the null situation as well as the critical values to determine aberrance of given statistical properties are unknown. Several authors have provided simulation studies of the 3PL model with aberrants modelled by spuriously low and high modifications of IRT-generated responses which can be considered as simulations of random cheating or random guessing,

respectively (e.g., Levine & Drasgow, 1982, 1983; Levine & Rubin, 1979). These authors have found that the $LR/\bar{\sigma}^2$ and L_0 indices yield similar results in detecting rates of such aberrants. Thus, in detecting these kinds of aberrance, the Gaussian model is not more powerful than the 3PL model (however, from theoretical considerations more power was expected, see Levine and Drasgow (1983), p.119).

Resuming, the L_0, ZL_0, LR and $\bar{\sigma}^2$ indices can be calculated for the 1PL and 2PL models, too (the L_0 and ZL_0 indices are included in Hoijtink, 1986, study). Some extensions of these indices are relatively easy to obtain. So, in order to obtain extensions of LR or $\bar{\sigma}^2$, more realistic variable ability models can be considered. These extensions can be employed to test the null hypothesis of constant ability against different alternative hypotheses of a person's ability distribution. However, each extension should be checked to see whether it is really more powerful for detecting the specific kind of aberrance in question. The likelihood-based analogue of UW can be obtained as a version of L_0 defined on the subset of items and can be used to measure the within-subset person's fit. Drasgow, Levine and Williams (1985) have considered the generalization of ZL_0 for a multiple-choice test with omitted items. From their simulation study it follows that the generalized ZL_0 index is clearly superior to the original ZL_0 index in detecting spuriously low modifications of IRT-generated responses. Unfortunately, this may not be said in the case of detecting spuriously high modifications.

Quite recently, particularly for the RM, Molenaar (1986) has proposed yet another index defined as the conditional probability of a person's pattern $\underline{X}$ given the test score group r to which this persons belongs, i.e.,

$$P(\underline{X}|r) = \frac{1}{\gamma_r} \prod_{i=1}^{L} \{\exp(-b_i)\}^{x_{vi}},$$

where b_i is the difficulty of item i and γ_r is the elementary symmetric function of order r (Fischer, 1974). If this probability has a very low value, its pattern has to be classified as aberrant. For the RM, this index is justified because the total test score is a sufficient statistic for the ability parameter. This is not the case for the 2PL and the 3PL model and for these models the considered index is thus not appropriate. The purpose to built this index was to eliminate the test-score-group dependence as was proved by Hoijtink (1986a) for all known indices in the case of the RM. For a more extensive discussion of this new index, see Molenaar (1986).

Extended Caution Indices

Several authors have offered indices to detect aberrants which are based on classical summary statistics (e.g., the correlation between a person's pattern and the vector of proportions correct or the Guttman vector). For a review of these indices, Harnisch and Linn's (1981) paper is recommended. Tatsuoka and her associates

have proposed a link between the classical and the IRT approach (Tatsuoka, 1984; Tatsuoka & Linn, 1983). In Tatsuoka's method, a person's pattern of responses is determined as aberrant if the pattern is not correlated (to some extent) with the vector of group or individual probabilities correct. The indices in this method are direct extensions of Sato's caution index and his Student-Problem approach, and are therefore called extended caution indices.

If P_{vi} denotes the probability of a correct response given by person v to item i in an IRT model, then

$$\underline{P}_v = (P_{vi}, P_{v2}, \ldots, P_{vL})$$

is the vector of individual probabilities correct and

$$\underline{G} = (G_1, G_2, \ldots, G_L), \qquad \text{where } G_1 = \frac{1}{n} \sum_{v=1}^{n} P_{vi} ,$$

is the vector of group probabilities correct. The extended caution indices are defined as the complement of the ratio of two covariances (or correlations) between different combinations of vectors: $\underline{X}$, $\underline{P}_v$, $\underline{G}$ and the usual vector of observed proportions correct (Tatsuoka, 1984; Tatsuoka & Linn, 1983). Two of them are recognized as representative in this family, namely ECI2 and ECI4, defined as

$$ECI2 = 1 - \frac{Cov(\underline{X}, \underline{G})}{Cov(\underline{P}_v, \underline{G})}$$

and

$$ECI4 = 1 - \frac{Cov(\underline{X},\underline{P}_v)}{Cov(\underline{G},\underline{P}_v)} .$$

Note that the denominator is the same in both indices and can be considered as a normalizing constant. Thus, ECI2 measures the similarity of an observed pattern to a group-normed variable and ECI4 measures the similarity of an observed pattern to a individual-normed variable (the ECI2 index is more group dependent whereas ECI4 more individual oriented). If a value of the index is far from zero, tne person can be determined as aberrant. Since the indices are functions of ability (see Tatsuoka, 1984, for calculations of the conditional expectations and variances of indices for the RM), the author has proposed their standardized versions ZECIm as

$$ZECIm = \frac{ECIm - E(ECIm|\theta)}{Var^{1/2}(ECIm|\theta)} , \quad \text{where } m = 2,4.$$

Tatsuoka has also shown that the Guttman's vector (a vector whose first elements are one's and the remaining are zero's) at each ability level has the smallest ZECI2 value while the reversed Guttman's vector has the largest ZECI2. This property is useful in the interpretation of ZECI2 and is a realization of Sato's purposes.

The distributions of the ZECI2 and ZECI4 indices in the null situation are still not sufficiently known. In Tatsuoka's (1984) opinion, for logistic models, tne distributions are approximately

standard normal. So, a fixed percentile of the standard normal distribution nas been used as a critical value to detect aberrants. For the RM, sample and test-score-group dependence of percentiles of these indices are illustrated by Hoijtink (1986a).

Finally, it should be noted that all extended caution indices are more or less group dependent. This is why they do not directly measure the fit of the person's pattern to the IRT model used. However, in contrast with other IRT indices, the ZECI2 allows us to compare a given pattern to the group-normed one, and so, can be utilized in student-class comparisons or to control for group effects (Harnisch & Linn, 1981). In addition, the ZECI4 index is very helpful to detect and diagnose students who consistently use an erroneous rule of operation (e.g., in aritnmetic problem tests). Applications with this aim and their computerization have been the subject of extensive study of several authors (for a review of methods and applications, see Tatsuoka, 1985).

Comparison of Indices in the Case of the Rasch Model

Several authors have performed a comparison of the indices, using both real and simulated data, in order to find a certain relationsnip between them (e.g., Birenbaum, 1985 ; Harnisch & Tatsuoka, 1983) and to evaluate their relative usefulness in detecting aberrants (e.g., Levine & Drasgow, 1983; Rudner, 1983). Unfortunately, these comparisons are too incomplete and too

inconclusive to be in a position to draw general conclusions. However, the authors have confirmed previous conclusions concerning less ability dependence of standardized indices in relation to their unstandardized origins. Only Hoijtink's (1986a) recent simulation study provides important results concerning certain dependence of indices distribution and its percentiles which, as a consequence, allow us to detect aberrants better (see below).

To be able to make more conclusive comparisons, all indices discussed in this paper have been divided into three groups in regard to what they measure: indices of the overall person's fit (UT/ZUT, WT/ZWT, L_0/ZL_0, $P(\underline{X}|r)$); indices of a more specific kind of person's fit (UB, UW, χ^2, $LR/\hat{\sigma}^2$) and indices of the unsimilarity of the person's pattern to the group- (ECI2/ZECI2) or individual-normed one (ECI4/ZECI4). Indices of the first group measure the same and can be compared each other. The indices of remaining two groups are more specifically oriented, thus, the supremacy one of these indices can be considered only in terms which of them better fulfils needs in question.

In the case of the first group of indices, residuals are accumulated over all person's responses (with the UT/ZUT, WT/ZWT indices) or the unexpectness of the whole pattern (with the L_0/ZL_0, $P(\underline{X}|r)$ indices) is measured. Therefore, these indices can be utilized to test the overall goodness of a person's fit. So, as is usual in a procedure of testing a statistical hypothesis, a value of the index larger (lower) than some critical value allows us to classify the person as aberrant. The procedure does not exclude possibly wrong

decisions, i.e. the declaration of the IRT-behaving person as aberrant (the type I error). The probability of the type I error will be independent of the sampled item and person parameters if the critical value is equal to the fixed percentile of the proposed index distribution in the null situation. However, Hoijtink (1986a) has proved that in the RM a fixed percentile of the index distribution is sample- and test-score-group dependent. The sample dependence means that the overall percentile differs as a function of the item difficulty and ability distributions. (e.g., the estimates of the overall 95-th percentile of ZL_0 are significantly different if the difficulty distribution is uniform on the interval $[-3,3]$ and the ability distribution is standard normal or uniform on $[-3,3]$). The test-score-group dependence means that the percentile differs over test score groups (e.g., the variance of ZL_0 is much less for both extremely low and high test score groups than for medium test score groups). Thus, from Hoijtink's study it follows also that the critical value to identify aberrants may not be determined over all score groups, but have to be calculated for each score group separately. Only then we are in a position to get a fixed type I error.

To preclude the test-score-group dependence, Molenaar (1986) has introduced the $P(\underline{X}|r)$ index. This index is justified theoretically for the RM and, so far, is the only one which is no function of the ability estimate. If the test-score-group dependence of this index were confirmed, the $P(\underline{X}|r)$ index would be the most appropriate index for the RM. Alternatively, the L_0 index as an index relatively simple to calculate and to study theoretically has to be consid-

ered. For the L_0 index, the approximation of its appropriate critical values is not exclusive. In conclusion, at the moment the $P(\underline{X}|r)$ and L_0 indices are suggested for use in a Test Service System to detect aberrants. Until now, there is no theoretical method to calculate or approximate these critical values, they have to be approximated by a simulation of the null situation. Another problem is that observed data almost always include aberrants, thus practically the null situation is unknown. So, only an iterative strategy of which each iteration consists of an approximation of the appropriate critical values and a procedure to detect and remove aberrants from data, can be helpful. A procedure with the algorithm given below realizes this iterative strategy and can be used to detect aberrants in tests in which the items are already of Rasch's structure.The procedure is as follow:

(1) estimate item and person parameters for the data set in question;

(2) simulate item responses with difficulties from (1) for a sufficiently large number of persons with abilities from (1), generated in accordance with the RM;

(3) calculate the index values for persons simulated in (2) and collect them per test score group;

(4) calculate critical values per test score group, (i.e., calculate a fixed percentile for the subgroups in (3));

(5) indicate and remove aberrants in the data set from (1) per test score group, with the help of the critical values obtained in (4);

(6) a) turn back to step (1) if still aberrants are indicated in (5);

b) end the procedure if no aberrants are detected in (5) anymore.

The procedure starts with an observed data set, and ends with a data set without aberrants. Note that the estimates of the last iteration are final estimates of item and person parameters, free of the influence of aberrants. If the $P(\underline{x}|r)$ index is test-score-group independent, an overall fixed percentile in step (4) can be calculated, and thus aberrants can be indicated in step (5) on the basis of a common critical value. This would save a lot of calculation time and would increase the accuracy of detection of aberrants. The above described procedure is a simplified version of Hoijtink's (1986b) more general procedure. Hoijtink's procedure can be used if some items in a test do not satisfy the RM assumptions. In his procedure aberrants are removed in the first turn and non-fitting items in the second turn.

The indices of the second group, i.e., UB, UW, χ^2, LR and $\bar{\sigma}^2$, can be used to test a specific form of person's fit. Namely, LR and $\bar{\sigma}^2$ are created especially to test persons for the presence of normally distributed ability fluctuations (i.e., to test the null hypothesis that ability is constant during the test against the alternative hypothesis that ability is a normally distributed variable). The UB and χ^2 indices can be employed to test the null hypothesis that the person's ability is the same in each group of items. Testing this hypothesis is of importance in order to be able

to diagnose specific item-person interaction or item bias. Diagnoses of certain item-order-person interactions are also possible; however, with less power because an appropriate specification of subsets of items is usually discussable. The UW index, as a measure of the within-subset-of-items person's fit, seems to be helpful to explain causes of the overall person's nonfit, though it is less suitable to test a general hypothesis about the person's ability. Because all these indices are more specifically oriented thus, selecting the appropriate index from them depends on the specific needs in question. Since for the TSS users, testing for the presence of specific item-person interaction is very useful, the UB and χ^2 indices and the following procedure is recommended:

formalize a fixed popular specific item-person interaction by specifying appropriate subsets of items;

utilize steps (2), (3) and (4) from previously described procedure using the UB or χ^2 index to calculate the appropriate critical values;

calculate the index value for all persons or only for aberrants already detected with the help of the overall fit index (e.g., $P(\underline{X}|r)$);

indicate per test score group persons for which index value is larger than the critical value as persons of chosen item-person interaction.

Optionally, for aberrants already detected, a specific interaction can be tested (the user defines appropriate subsets of items). For

this purpose, a properly presented graph of standardized person's residuals could be also very helpful for the TSS user to formulate a hypothesis about the specific item-person interaction.

The indices of the third group (ECI2/ZECI2, ECI4/ZECI4) are more or less group dependent; thus they do not measure the fit of the person's pattern to an IRT model directly. However, in contrast with all other indices, they allow us to compare a given pattern to another one. Therefore, the extended caution indices can be applied in a TSS when it is necessary to make comparisons within a student-class or a student-reference group, or to analyse the effects of training, interventions, or policies on educational attainments of different groups.

Finally, several authors have proposed robust estimation of ability in IRT models (e.g., Mislevy & Bock, 1982; Wainer, 1983; Wainer & Wright, 1980). These procedures are corrections to the usual IRT ability estimation procedure and to its standard error for aberrant persons. The corrections consist of weighting the unexpectness of the person's responses. However, Smith (1985) has shown that, for the RM, recalculated maximum-likelihood estimates of ability recover simulated abilities better than robust estimates. In these recalculations, aberrant responses detected in an earlier estimation are replaced by missing values. Another procedure for better item and person estimation in the RM is proposed by Choppin (1985). In this procedure only guessed responses are replaced by missing values. Smith's and Choppin's procedures can be considered as alternative approach to the problems appearing when

aberrants are present in a sample; they can also be utilized to get better item and person parameter estimates, free from the influence of aberrants.

Discussion

At the moment, some progress is observed in detecting persons with aberrant response patterns in the case of the Rasch model. From simulation studies for this model it follows that critical values to declare aberrance have to be determined for each score group separately instead of as is the usual practice, over all score groups simultaneously. Thus, using only different appropriate critical values calculated per test score group, the probability of misclassification of a Rasch-behaving person as aberrant can be constant and under control. A new index to identify aberrants is introduced which is theoretically justified for the Rasch model. Finally, a procedure to detect and remove aberrants and to exclude items which do not satisfy assumptions of the Rasch model is proposed. Now, more effective detection of aberrants and, consequencently, better item and person parameter estimation can be expected.

References.

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. Educational and Psychological Measurement, 45, 523-534.

Choppin, B.H.L. (1983). A two-parameter latent trait model. Evaluation in Education, 9, 43-62.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Drasgow, F., & Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. Applied Psychological Measurement, 10, 59-67.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Fischer, G.H. (1974). Einfuhrung in die Theorie Psychologicher Tests. Bern:Huber.

Harnisch, D.L. (1983). Item response patterns: applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Harnisch, D.L., & Tatsuoka, K.K. (1983). A comparison of appropriateness indices based on item response theory. In R.K. Hambleton (Ed.), Applications of item response theory. Vancover, B.C.: Educational Reseach Institute of British Colombia.

Hoijtink, H. (1986a). Detecting aberrant response patterns in the unidimensional scaling model of Rasch. Heymans Bulletin, HB-86-792-SW. Groningen: R.U.Groningen.

Hoijtink, H. (1986b). Rasch schaal constructie en person fit. (Technical Report). Groningen: Psychologiscne Instituten, R.U. Groningen.

Kingsbury, G.G. (1985). The identification of unusual response patterns. (Technical Report). Portland Public Schools.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.

Levine, M.V., & Drasgow, F. (1983). Appropriateness measurement: validating studies and variable ability models. In D.J. Weiss (Ed.) New horizons in testing: latent trait theory and computerized adaptive testing. New York: Academic Press.

Levine, M.V., & Drasgow, F. (1984). Performance envelopes and optimal appropriateness measurement. (Technical Report MS 84-5). University of Illinois.

Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test score. Journal of Educational Statistics, 4, 269-290.

Ludlow, L.H. (1985). A strategy for the graphical representation of Rasch model residuals. Educational and Psychological Measurement, 45, 851-859.

Mislevy, R., & Bock,R.D. (1982). Biweight estimates of latent ability. Educational and Psychological Measurement, 42, 725-737.

Molenaar, I.W. (1986). Person fit indices in the Rasch model. (forthcoming).

Rudner, L.M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Smith, R.M. (1985). A comparision of Rasch person analysis and robust estimators. Educational and Psychological Measurement, 45, 433-444.

Smith, R.M. (1986). Person fit in the Rasch model. Educational and Psychological Measurement, 46, 359-372.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. Journal of Educational Statistics, 10, 55-73.

Tatsuoka, K.K., & Linn, R.L. (1983). Indices for detecting unusal patterns: links between two general approaches and potential applications. Applied Psychological Measurement, 7, 81-96.

Trabin, E.T., & Weiss, D.J. (1983). The person response curve: fit of individuals to item response theory models. In D.J. Weiss (Ed.) New horizonts in testing: latent trait theory and computerized adaptive testing. New York: Academic Press.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D.J. Weiss (Ed.) New horizonts in testing: latent trait theory and computerized adaptive testing. New York: Academic Press.

Wainer, H., & Wright, B.D. (1980). Robust estimation of ability in the Rasch model. Psychometrika, 45, 373-391.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, N.J.: Educational Testing Service.

Wright, B.D. (1977). Solving measurement problems with Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B.D., & Stone, M.H. (1979). Best test design. Chicago: MESA Press.

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis. Chicago: MESA Press.

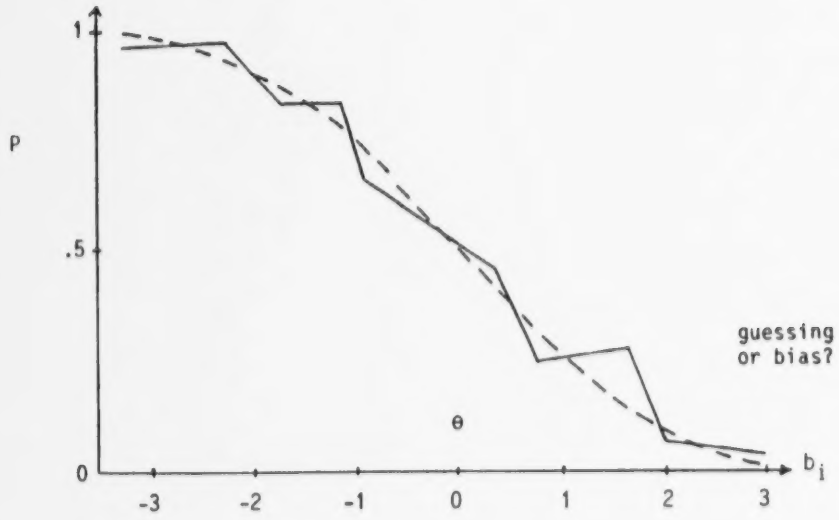

Figure 1. Observed and expected person response curve.

(solid line, observed PRC; dashed line, expected PRC)

Author's Note

The author is indebted to Ron J.H. Engelen and Wim J. van der Linden for their comments on an earlier version of the paper. The content of the paper is however fully his own responsibility.

Titels van Recente Rapporten

R-86-1 C. Sluijter, Het bijwerken van een item bank

R-86-2 D.L. Knol, Inventarisatie van automatische itemselectie procedures voor Raschschalen

R-86-3 G.R. Buning, T.J.H.M. Eggen, H. Kelderman & W.J. van der Linden (red.), Het gebruik van het Raschmodel voor een decentraal toetsservicesysteem

R-86-4 J. Kogut, Review of IRT-based indices for detecting and diagnosing aberrant response patterns

R-86-5 D.L. Knol, Een overzicht van meerdimensionale itemresponsmodellen

Rapporten in deze reeks kunnen tegen kostprijs besteld worden bij: Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, Postbus 217, 7500 AE Enschede, tel. 053-893588

Titles of Recent Research Reports

RR-86-1 W.J. van der Linden, The use of test scores for classification decisions with threshold utility

RR-86-2 H. Kelderman, Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups

RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by zero-one programming

RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one programming approach to Gulliksen's random matched subtests method

RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity analysis with k sets of variables: An altenating least squares method with optimal scaling features

RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes approach to item banking

RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of parallel tests by zero-one programming

RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models for paired comparisons with ties

Research Reports can be obtained at costs from Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

Een publicatie
van de faculteit der
Toegepast Onderwijskunde
van de Universiteit Twente,
Postbus 217,
7500 AE Enschede

Toegepaste Onderwijskunde